

Keil, Stefan; Böhm, Peter; Rittberger, Marc

Qualitative web analytics. New insights into navigation analysis and user behavior - a case study of the German Education Server

Pehar, Franjo [Hrsg.]; Schlögl, Christian [Hrsg.]; Wolff, Christian [Hrsg.]: Re:inventing Information Science in the networked society. Glückstadt : Hülsbusch 2015, S. 252-263. - (Schriften zur Informationswissenschaft; 66), 10.5281/zenodo.17938



Quellenangabe/ Reference:

Keil, Stefan; Böhm, Peter; Rittberger, Marc: Qualitative web analytics. New insights into navigation analysis and user behavior - a case study of the German Education Server - In: Pehar, Franjo [Hrsg.]; Schlögl, Christian [Hrsg.]; Wolff, Christian [Hrsg.]: Re:inventing Information Science in the networked society. Glückstadt : Hülsbusch 2015, S. 252-263 - URN: urn:nbn:de:0111-dipfdocs-189486 - DOI: 10.25657/02:18948

<https://nbn-resolving.org/urn:nbn:de:0111-dipfdocs-189486>

<https://doi.org/10.25657/02:18948>

Nutzungsbedingungen

Dieses Dokument steht unter folgender Creative Commons-Lizenz: <http://creativecommons.org/licenses/by/4.0/deed.de> - Sie dürfen das Werk bzw. den Inhalt vervielfältigen, verbreiten und öffentlich zugänglich machen sowie Abwandlungen und Bearbeitungen des Werkes bzw. Inhaltes anfertigen, solange Sie den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

This document is published under following Creative Commons-License: <http://creativecommons.org/licenses/by/4.0/deed.en> - You may copy, distribute and render this document accessible, make adaptations of this work or its contents accessible to the public as long as you attribute the work in the manner specified by the author or licensor.

By using this particular document, you accept the above-stated conditions of use.



Kontakt / Contact:

DIPF | Leibniz-Institut für
Bildungsforschung und Bildungsinformation
Frankfurter Forschungsbibliothek
publikationen@dipf.de
www.dipfdocs.de

Mitglied der

Leibniz
Leibniz-Gemeinschaft

Qualitative Web Analytics: New Insights into Navigation Analysis and User Behavior

A Case Study of the German Education Server

Stefan Keil, Peter Böhm, Marc Rittberger

German Institute for International Educational Research
Schloßstraße 29, 60486 Frankfurt am Main, Germany
{keil, boehm, rittberger}@dipf.de

Abstract

Web analytics is a common approach to monitoring and analyzing user behavior on websites. We investigated three different research aspects which can be addressed using web analytics data. Our main concern was the data quality, followed by general findings on user behavior as well as potential usability issues. We chose an iterative, qualitative approach in order to address all three aspects. Further, we annotated usage data in detail to achieve a deeper understanding of possible user intentions. As a result, we conclude that the use of web analytics data, captured with a modern and widely used tool, bears some limitations for usability analysis, as semantic problems occur that are often overlooked. Further pre-processing is needed to reconstruct the real clickstream when an analysis of the navigation and user behavior is planned. Some hints at usability issues could be found by detecting movement patterns between certain page types.

Keywords: Web analytics, User behavior, Qualitative analysis, Case study

In: F. Pehar/C. Schlögl/C. Wolff (Eds.). Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science (ISI 2015), Zadar, Croatia, 19th–21st May 2015. Glückstadt: Verlag Werner Hülsbusch, pp. 252–263.

1 Introduction

Nowadays, web analytic tools are widely used especially since easy-to-implement solutions like Google Analytics have been available for free. However, even powerful tools provide only highly aggregated reports and quantitative metrics. The more complex a website is, the more important it is to analyze and understand the usage and navigation behavior of its users.

From a scientific point of view, web analytics has been found to be of high practical relevance but not widely researched. In particular, this holds true for modern web analytic approaches like JavaScript-based data capturing.

Bringing together both aforementioned issues, it seems important to introduce qualitative scientific approaches to modern web analytic methods in order to identify user behavior and other issues concerning large websites that might not be identified when using only quantitative standard metrics.

So we conducted a qualitative study with the goal to overcome disadvantages occurring when working exclusively with quantitative methods and to provide deeper insights through annotating elaborated navigation patterns and possible navigation or usability issues.

In the following, we briefly discuss the concept of navigation and user behavior plus web analytics as a method of investigation. We focus on problems in data capturing and data quality. The conducted study (section 4) deals with these aspects in the context of the German Education Server (GES), a specialized web catalogue in the domain of Education. After discussing the results (section 5), an outlook for further research is given.

2 Web analytics

We see web analytics as part of web monitoring, which Höchstötter & Lewandowski (2014: 28) define as a time-critical and systematic collection and analysis of data on the web. Web analytics primarily refers to the on-site analysis which is mainly focused on actions performed on the analyzed website (e.g. page views, downloads). Complementary off-site analysis focuses on aspects related yet remote to the analyzed website and comprises for example (back-) link analysis or social media mentions. On-Site analysis is also

known as clickstream analysis (e.g. Kaushik 2010), the clickstream describing the “stream” of page views and other actions (e.g. searches or self-defined events) that a user performs during a session. In a broader sense it can also be referred to as transaction log analysis (e.g. Sheble & Wildemuth 2009). Traditionally, quantitative analyses are highly predominant in the field of web analytics.

2.1 Data capturing

In general, two different approaches for data capturing can be distinguished in the analysis of websites. First, (server) log files can be filtered and used for data analysis. Such log files are the oldest data source available for web analytics, since such logs are created by web servers by default. Log files have been more and more replaced by a second approach for data capturing, a so-called page tagging method based on JavaScript and Cookies. This method enables (selective) capturing of page views and other (customized) events. Compared to log files, page-tagging has some advantages for analyzing user behavior.¹ For example, caching which was estimated to absorb about 45% of all transactions between a client and a server² does not apply to the page tagging technique. Another disadvantage of server log files in this context is that every single file request is captured, including those for image and style sheet files. This makes filtering for page views inevitable. Such filtering is not required for the page tagging method since only page views are captured by default.

The advent of the page tagging method has also popularized the use of the software-as-a-service (SaaS) approach. Here, data capturing and analysis are entirely delegated to service providers. Common products like Google Analytics are available free of charge. Piwik is an open source alternative to Google Analytics and it allows hosting on one’s own server. Self-hosting is more suitable regarding data privacy issues and grants full control over the captured data. Furthermore, full access to the raw data also allows more detailed analyses.

¹ The following list of arguments is not comprehensive. For a more complete overview of advantages and disadvantages of page tagging and log files, see Wikipedia (2015a).

² Nicholas (2000) according to Sheble & Wildemuth (2009: 168)

2.2 Error types and problems

Scientific literature on web analytics mostly refers to log files as a data source. Problems concerning the page tagging method are mainly an object of interest in blogs and how-to manuals (e.g. Kaushik 2010).

Regardless of the data source, two groups of possible problems can be distinguished: syntactical and semantic errors.

Syntactical problems can usually be ascribed to ascertainment errors which often result in a lack of data. In log files, page views may be missing due to caching. With page-tagging, deactivated JavaScript and Cookies in the users' web browsers may result in missing records. So-called POST forms present a problem that occurs in both approaches. With this type of web forms, the form data is transmitted in a separate data stream from the web browser to the web server. This means that the field values do not appear in the URL of the target page of the form, as would be the case with GET forms. Depending on the size and content of a web form, the decision for a POST or a GET form is often arbitrary.

Semantic errors result in incomprehensive clickstream data. One cause of semantic errors are the afore-mentioned syntactical errors. Missing clickstream data can lead to incomprehensive sessions. Another source of semantic errors are re-entries to the website without a change of the referrer. For example, if two direct entries are performed by one user in a certain time period assumed to be one session, it is possible that no link connection exists between the single pages. This can occur when using direct links which are pasted into the browser's address bar or when using RSS feeds. However, the main source of semantic errors is tabbing. This phenomenon occurs when a user opens a link within the visited website in a new browser tab (or browser window), than uses this new tab for further browsing, before closing it and continuing to browse in the previous tab. The act of tab switching cannot be captured by web analytic tools. Instead, they normalize clickstream actions according to their timestamps.³ When tabbing occurs in a session, this normalization leads to mingled session data that suggests a browsing history which never happened. Common web analytic tools like Google Analytics and Piwik do not account for tabbing.

³ See e.g. Kaushik (2010: 50).

3 Navigation and user behavior

Traditional web analytics mostly uses relatively simple metrics like the number of page views, or the time spent on the website during a visit. Not all of these metrics are directly affected by the semantic errors described in the previous section. However, most metrics and approaches dealing with navigation sequences and user behavior are affected. These more sophisticated but rarely used approaches are introduced in this section.

The concept of lostness is a metric that takes the user's route through a website into account and refers to the feeling of being disorientated or lost in a hypertext system. Basically coined by Smith (1996), Otter & Johnson (2000) enhanced the concept. The link weighted lostness metric respects different types of links in a hypertext system in terms of their likelihood to cause lostness. The measure is based on the number of nodes (i.e. page views) which are required to fulfill a retrieval task. Each difference (i.e. additional page views) increases the lostness and thereby indicates that the user diverged from the optimal navigation path. Otter & Johnson's studies show that it is possible to measure a small correlation between the calculated lostness and self-reported lostness as well as between the calculated lostness and the total task time even if it is not significant.

Ahrens et al. (2014) argue that log file analysis can complement usability tests. The authors developed a plugin for Piwik to analyze navigational trails in order to identify usability problems. They analyzed the frequency of click path subsequences with regard to hierarchical movements on a website, showing for example that the start page is viewed comparably seldom in the most frequent subsequences.

Canter et al. (1985) established a variety of patterns which can occur when users navigate through database environments. A path, a ring, a loop and a spike are described as basic patterns (Canter et al. 1985: 95 f.). Transferred to modern web analytics, a path can be defined as a sequence of page views in which each page is viewed only once. A ring is a sequence where the first page and the last page are the same. By contrast, a loop which basically matches a ring can contain multiple rings in itself. A spike is a sequence of page views which ends at a certain point and is then followed back in reverse order. Furthermore, Canter et al. (1985: 100) define more specialized and complex search strategies. These complex search patterns require a high number of actions per session to be possibly identified.

Compared to standard metrics, it can be assumed that these more sophisticated metrics introduced in this section would yield additional insights regarding possible usability and navigation problems of a web portal. To verify this assumption, we conducted a study to identify the patterns described by Canter et al. (1985) in order to examine if their occurrence is linked to certain usability issues or tendencies in and outcomes of a session.

4 Our study

Our study is based on data captured using the self-hosted web analytic tool Piwik⁴. Three aspects were examined: data quality, user behavior and potential insights into usability of the investigated website. We placed a main focus on data quality in consideration of the user behavior. Selected sessions were annotated according to the web navigation patterns by Canter et al. (1985) to see if any connection between their occurrence and a potential success of a session can be established.

In the remainder of this section, the examined website is introduced followed by a description of the data pre-processing and selection as well as a delineation of the method.

4.1 The German Education Server as the object of study

The German Education Server⁵ (GES) is the largest web portal in the domain of German educational information. It is focused on providing high-quality metadata about specialized and relevant information in education and educational research which is distributed across the web (cf. Kühnlenz et al. 2012: 23). Furthermore, the GES is a very heterogeneous system with a variety of different databases, catalogues and dossiers. Such databases include e.g. institutions and competitions. Its main part is a specialized web catalogue⁶ with ten basic categories which are further specialized in up to eleven hierarchi-

4 We used Piwik as it was the only JavaScript-based web analytics tool in use at the German Education Server granting access to its database of detailed session data.

5 www.bildungsserver.de

6 See Griesbaum et al. (2009: 22).

cally and poly-dimensionally organized sub-sections. This catalogue consists of edited web pages on which selected entries from the aforementioned databases are clustered thematically. A quantitative study of the user behavior on the GES was conducted by Böhm (2010) showing some weak dependencies between referrer types and standard web analytic metrics.

4.2 Basic set and cleansing

Our basic set contains 232,559 sessions from January 2014 which were carried out by 199,620 unique users of the GES. During data cleansing routines (e.g. checking for the right data format and allowed values), a syntactical error in the Piwik data was discovered.⁷ This error resulted in incomplete session data and required the exclusion of 12,497 affected sessions, containing 105,656 actions.

Grouped by referrer type, the sessions are distributed as follows: 30,540 (13.8%) direct entries, 158,012 (71.8%) search engine entries, 29,502 (13.4%) website entries and 2,008 (0.9%) campaign entries. An excess proportion of search engine entries can thus be observed.

To ensure the annotation of all patterns described by Canter et al. (1985), a quorum of six actions per sessions was set. This quorum equates to the number of actions needed to perform a loop. Furthermore, it is hardly possible to interpret a session (e.g. to capture a notion of the user's information need) with fewer than six actions.

Due to limitations of our methodological approach, the maximal number of actions per session was limited to 25. Considering their frequency, more than 25 actions cannot be characterized as typical usage of the GES. Leaving 25,710 sessions in the pool, we randomly picked a sample of 25 sessions. To ensure that at least three sessions per referrer type were included, a small oversampling of campaign referrals was allowed for. The final sample consisted of four sessions with a direct entry, three with a website entry, three with a campaign and 15 with a search engine entry.

⁷ Some actions have been captured but no URL has been assigned. Hence only the fact that some action has been carried out but not its actual URL was available from the Piwik database.

4.3 Method

We used a qualitative and iterative approach for our analysis. All sessions were manually reconstructed using the actual GES website and during this process supplemented and annotated with additional information about the links and navigation elements used as well as the pages' contents, topics and displayed page elements.

All sessions were formatted in tables (see table 1) to provide a consistent overview. This overview was further enhanced by the creation of graphs representing the steps of each session (e.g. nodes #1 to #9 in fig. 1). The graphs provided more insights concerning the patterns described by Canter et al. (1985) (see fig. 1).

Table 1. Example of the table structure

navigation element	Used element type (e.g. left sidebar, content)	
Action #	page topic	
	page URL	
	page type	time spent

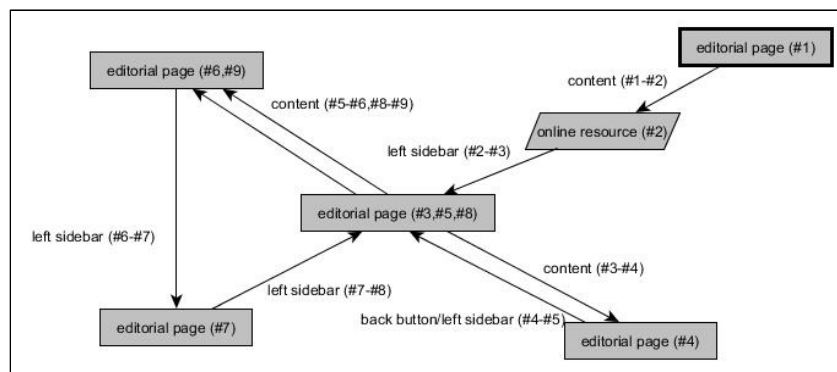


Figure 1. Graph illustrating an analyzed session

Furthermore, each session was described in natural language, taking more information like the topic of a page and its content into account.

At last, a detailed discussion and interpretation of the session was given, with a focus on the following aspects:

- Traceability (to what degree could the clickstream data be reproduced on the actual website),

- Potential success of the session if an information need was recognizable,
- General behavior (e.g. idiosyncrasies regarding the use of different page types),
- Usability,
- Particularities concerning the recognized patterns.

As already mentioned, quantitative approaches are more common when dealing with web analytics data. Our qualitative approach was chosen in order to include all described aspects and dimensions and to gain an insight into possible errors and problems with the captured data which can be used to develop further, possibly algorithmic approaches to retrace clickstream data. Besides, usability is hardly measurable based on highly aggregated metrics given the high degree of heterogeneity of the web portal.

5 Results

We present our results structured by the three different research aspects introduced in section 4.

5.1 User behavior

The user behavior was mainly examined by annotating the sessions according to the patterns by Canter et al. (1985). In general, none of the patterns could be linked to a specific session outcome. Two particularities were identified. First, sessions with paths (occurring in seven sessions) tend to deal with more than one subject. Sessions with many paths or a very long path may either point to users who explore the website or to a usability problem as users cannot satisfy their information need. Second, in all analyzed sessions at least one spike could be observed.⁸ Taking the different page types into account which were accessed during a spike, a more differentiated picture emerges. As the GES mainly provides metadata about other websites, the main usage scenario is the search for specific information in the educational domain. Most GES pages present metadata about other websites in a list-like

⁸ The return of a user to a tab abandoned earlier was seen as a spike, even if no new page view was captured.

manner.⁹ As part of a search process, pages with detailed information about the resources as well as the resources themselves are viewed. Occasionally, the user returns to the result or editorial pages and continues the search. This pattern (a spike linking result or editorial pages with detail views and external sites) can be seen as part of a usual search scenario and relevance judgement process, confirming the intended usage of the system. Spikes that appear when database entry points and database result pages are linked can be interpreted as indicators of a problem. Users seem to find it difficult to phrase their information need so further support should be given regarding queries.

5.2 Usability insights for the GES

Taking the small sample size into account, the following findings should be seen as first indicators to be verified in future studies.

In one case, the movement between several editorial pages hints at a problem in the page titles and their anchor labels in the main menu. Titles and menu labels are in a hierarchical relationship to each other (“school” as an “institution” in a specific area of education) but were arranged at the same hierarchy level. It can be assumed that the user wants to return to an already known page but is not able to find the right path. In terms of Pirolli’s Information Foraging Theory, it can be assumed that the labels exude too similar information scents because they are semantically too closely related to each other.

As already mentioned, three sessions showed an increased occurrence of spikes between database entry points and result pages. Due to the POST method for web form data transmission, we do not know if and how the queries were altered but in general, indication of a missing support in query formulation can be assumed.

In two sessions, users showed a reluctant behavior on date-based search result pages compared to other result pages which were generated on content-based aspects. Users tended to spend less time on those date-based search result pages and they clicked on fewer search results. Additional content-based filter opportunities on the result page level might be a tool to support the users in completing their retrieval tasks.

⁹ This applies to result pages of the different databases as well as to editorial pages.

5.3 Data quality

Considering the data quality, the frequency of incorrect clickstreams was very high at 13 out of 25 sessions. In seven cases, a possible and plausible reconstruction of the clickstream could be established after taking tabbing into account. In six cases, tabbing was no suitable explanation. Furthermore, the semantic errors in the clickstream data were so severe that the actual navigation paths could not be retraced. Regarding typical data or information quality dimensions like completeness, validity, consistency, timeliness, accuracy or standards based (Wikipedia 2015b), we see incorrect clickstream data as an accuracy issue. The captured clickstream does not represent the navigation path the user really followed.

This outcome substantiates the need for a careful, possibly qualitative consideration of the data quality of clickstream data. For a larger-scale analysis, it would be necessary to preprocess the data to at least account for semantic errors due to tabbing.

6 Discussion and outlook

We presented a qualitative approach for analyzing usage data captured with web analytic methods. This approach yielded insights concerning data quality, user behavior and usability issues. The reconstruction and annotation of a sample of 25 sessions showed that more than half of the sessions had semantic errors and in seven cases a reconstruction was possible if tabbing was considered. This shows that clickstream data gathered using the page tagging method, which is generally considered to be superior to log files, can on close inspection be problematic. A further reconstruction is needed. This could be done by employing a crawler for link analysis to test for plausibility of the captured clickstream data, considering tabbing or refining the capturing settings. The refinement could be realized through parameters or customized variables for identifying used links and referring pages. Then, further investigation of the patterns and their validity concerning the indication of possible usability issues should be carried out. If evidence suggesting that sessions with spikes between certain page types are less successful holds true, a generalization for similar websites could be formulated.

References

- Ahrens, M.; Kölle, R.; Werner, K.; Mandl, T. (2014). Using Web Analytics to Investigate the Navigational Behavior of Users. In: Butz, A.; Koch, M.; Schlichter, J. (Eds.). *Mensch & Computer 2014 – Tagungsband*. Berlin: De Gruyter Oldenbourg, pp. 95–104.
- Böhm, P. (2010). *Ermittlung von Nutzungsweisen auf dem Deutschen Bildungsserver mittels Webanalyse-Verfahren*. Master thesis: Hochschule Darmstadt. <http://www.dipfdocs.de/volltexte/2011/3800> <10.3.2015>.
- Canter, D.; Rivers, R.; Storrs, G. (1985). Characterizing user navigation through complex data structures. *Behaviour & Information Technology* 4 (2), 93–102.
- Griesbaum, J.; Bekavac, B.; Rittberger, M. (2009). Typologie der Suchdienste im Internet. In: Dirk Lewandowski (Ed.): *Handbuch Internet-Suchmaschinen: Nutzerorientierung in Wissenschaft und Praxis*. Heidelberg: Akademische Verlagsgesellschaft AKA, pp. 18–52.
- Höchstötter, N.; Lewandowski, D. (2014). Websuche und Webmonitoring. In: Höchstötter, N. (Ed.). *Handbuch Webmonitoring 1. Social Media und Website-monitoring*. Berlin: Akademische Verlagsgesellschaft AKA, pp. 23–46.
- Kaushik, A. (2010). *Web analytics 2.0. The art of online accountability & science of customer centricity*. Indianapolis: Wiley Publishing, Inc.
- Kühnlenz, A.; Martini, R.; Ophoven, B.; Bambey, D. (2012): Der Deutsche Bildungsserver – Internet-Ressourcen für Bildungspraxis, Bildungsverwaltung und Bildungsforschung. *Erziehungswissenschaft* 23 (44), 23–31.
- Nicholas, D. (2000). *Assessing Information Needs: Tools, Techniques and Concepts for the Internet Age* (2nd ed.). London: Europa Publications.
- Otter, M.; Johnson, H. (2000). Lost in hyperspace: metrics and mental models. *Interacting with Computers* 13 (1), 1–40.
- Sheble, L.; Wildemuth B. M. (2009). Transaction Logs. In: Wildemuth, B. M. (Ed.) (2009). *Applications of social research methods to questions in information and library science*. Westport, Conn: Libraries Unlimited.
- Smith, P. (1996). Towards a practical measure of hypertext usability. *Interaction with Computers* 8 (4), 365–381.
- Wikipedia (2015a). Web Analytics. http://en.wikipedia.org/w/index.php?title=Web_analytics&oldid=641601878 <10.3.2015>.
- Wikipedia (2015b). Data Quality. http://en.wikipedia.org/w/index.php?title=Data_quality&oldid=649628939 <10.3.2015>.